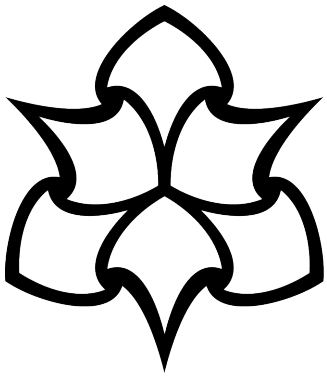

Expert Systems With Applications 168 (2021) 114351

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Credibility assessment of financial stock tweets

Lewis Evans [a,*], Majdi Owda [a], Keeley Crockett [a], Ana Fernandez Vilas [b]

[a] Department of Computing and Mathematics, Manchester Metropolitan University M1 5GD UK Manchester, England
[b] Ana Fernandez Vilas, I&C Lab, AtlantTIC Research Centre, University of Vigo, 36310 Pontevedra, Spain

ARTICLE INFO

Keywords:
Machine learning
Supervised learning
Twitter
Financial stock market
Feature selection

ABSTRACT

Social media plays an important role in facilitating conversations and news dissemination. Specifically, Twitter has recently seen use by investors to facilitate discussions surrounding stock exchange-listed companies. Investors depend on timely, credible information being made available in order to make well-informed investment decisions, with credibility being defined as the believability of information. Much work has been done on assessing credibility on Twitter in domains such as politics and natural disaster events, but the work on assessing the credibility of financial statements is scant within the literature. Investments made on apocryphal information could hamper efforts of social media's aim of providing a transparent arena for sharing news and encouraging discussion of stock market events. This paper presents a novel methodology to assess the credibility of financial stock market tweets, which is evaluated by conducting an experiment using tweets pertaining to companies listed on the London Stock Exchange. Three sets of traditional machine learning classifiers (using three different feature sets) are trained using an annotated dataset. We highlight the importance of considering features specific to the domain in which credibility needs to be assessed for – in the case of this paper, financial features. In total, after discarding non-informative features, 34 general features are combined with over 15 novel financial features for training classifiers. Results show that classifiers trained on both general and financial features can yield improved performance than classifiers trained on general features alone, with Random Forest being the top performer, although the Random Forest model requires more features (37) than that of other classifiers (such as K-Nearest Neighbours – 9) to achieve such performance.

1. Introduction

Investments made on stock markets depend on timely and credible information being made available to investors. Twitter has seen increased use in recent years as a means of sharing information relating to companies listed on stock exchanges (Ranco et al., 2015). The time-critical nature of investing means that investors need to be confident that the news they are consuming is credible and trustworthy. Credibility is generally defined as the believability of information (Sujoy Sikdar, Kang, O'donovan, Höllerer, & Adal, 2013), with social media credibility defined as the aspect of information credibility that can be assessed using only the information available in a social media platform (Castillo et al., 2011). People judge the credibility of general statements based on different constructs such as objectiveness, accuracy, timeliness and reliability (Sujoy Sikdar, Kang, O'donovan, & Höllerer, 2013). Specifically, in terms of Twitter, tweet content and metadata (referred to as features herein), such as the number of followers a user has, and how long they have been a member of Twitter have been seen as informative features for determining the credibility of both the content of the tweet, and the user posting it (de Marcellis-Warin et al., 2017). The problem with such features (namely a user's follower count) is that they can be artificially inflated, as users can obtain thousands of followers from Twitter follower markets within minutes (Stringhini et al., 2013), giving a false indication that the user has a large follower base and is credible (De Micheli & Stroppa, 2013). Determining the credibility of a tweet which is financial in nature becomes even more challenging due to the regulators and exchanges need to quickly curb the spread of misinformation surrounding stocks. Specifically, Twitter users seeking to capitalize on news surrounding stocks by leveraging Twitter's trademark fast information dissemination may be susceptible to rumours and acting upon incredible information within tweets (Da Cruz & De Filgueiras Gomes, 2013). Recent research has found that Twitter is becoming a hotbed for rumour propagation (Maddock et al., 2015). Although such rumours and speculation on Twitter can be informative, as this can reflect investor mood and outlook (Ceccarelli et al., 2016), this new age of financial media in which discussions take place on social media

* Corresponding author.
E-mail address: l.evans@mmu.ac.uk (L. Evans).

https://doi.org/10.1016/j.eswa.2020.114351
Received 16 March 2020; Received in revised form 28 August 2020; Accepted 17 November 2020
Available online 20 November 2020

demands mechanisms to assess the credibility of such posts. Repercussions for investors include being cajoled into investing based on apocryphal or incredible information and losing confidence in using a platform such as Twitter if such a platform can be used by perfidious individuals with impunity (De Franco et al., 2007). Twitter does not just act as a discussion board for the investor community, but also acts as an aggregator of financial information by companies and regulators. The financial investment community is currently bereft of ways to assess the credibility of financial stock tweets, as previous work in this field has focused primarily on specific areas such as politics and natural disaster events (Alrubaian et al., 2018).

To this end, one must define what constitutes a financial stock tweet and what is meant by determining the credibility of a financial stock tweet. This paper defines a financial stock tweet as any tweet which contains an occurrence of a stock exchange-listed company's ticker symbol, pre-fixed with a dollar symbol, referred to as a cashtag within the Twitter community. Twitter's cashtag mechanism has been utilised by several works for the purposes of collecting and analysing stock discussion (Oliveira et al., 2016, 2017; Cresci et al., 2018). Although tweets may be relating to a financial stock discussion and not contain a cashtag, this paper takes the stance that tweets are more likely to be related to stock discussions if cashtags are present, and this research focuses on such tweets. We define the credibility of a financial stock tweet as being three-fold: (1) is the cashtag(s) within the tweet related to a specific exchange-listed company? (2) how credible (based on the definition above) is the information within the tweet? and (3) how credible is the author circulating the information? We adopt the definition of user credibility from past research as being the user's perceived trustworthiness and expertise (Liu et al., 2012).

The main contribution of this paper is a novel methodology for assessing the credibility of financial stock tweets on Twitter. The methodology is based on feature extraction and selection according to the relevance of the different features according to an annotated training set. We propose a rich set of features divided into two groups – general features found in all tweets, regardless of subject matter, and financial features, which are engineered specifically to assess the credibility of financial stock tweets. We train three different sets of traditional machine learning classifiers, (1) trained on the *general* features, (2) trained on the *financial* features, and (3) trained on both general and financial feature sets – to ascertain if financial features provide added value in assessing the credibility of financial stock tweets. The methodology proposed in this paper is a generalizable approach which can be applied to any stock exchange, with a slight customisation of the financial features proposed depending on the stock exchange. An experiment utilising tweets pertaining to companies listed on the London Stock Exchange is presented in this paper to validate the proposed financial credibility methodology. The motivation of this paper is to highlight the importance of incorporating features from the domain in which one wishes to assess the credibility of tweets for. The novelty of this work lies in the incorporation of financial features for assessing the credibility of tweets relating to the discussion of stocks.

The research questions this paper will address are as follows:

RQ 1: Can features found in any tweet, regardless of subject matter (i.e. general features), provide an accurate measure for credibility classification of the tweet?

RQ 2: Can financial features, engineered with the intent of assessing the financial credibility of a stock tweet, provide improved classification performance (over the general features) when combined with the general features?

In addition to the methodology for assessing the financial credibility of stock tweets, the other key contributions of this paper can be summarised as follows:

- We present a novel set of financial features for the purpose of assessing the financial credibility of stock tweets
- We highlight the importance of performing feature selection for assessing financial credibility of stock tweets, particularly for machine learning models which do not have inherent feature selection mechanisms embedded within them.

The remainder of this paper is organised as follows: Section 2 explores the related work on the credibility of microblog posts. Section 3 provides an overview of the methodology used. Section 4 outlines the proposed features used to train the machine learning models. Section 5 describes the feature selection techniques used within the methodology. Section 6 outlines the experimental design used to validate the methodology. Section 7 provides a discussion of the results obtained. Section 8 concludes the work undertaken and outlines avenues of potential future work.

2. Background

Although there has been no research on the credibility of financial stock-related tweets, work does exist on the credibility of tweets in areas such as politics (Sujoy Sikdar, Kang, O'donovan, Höllerer, & Adal, 2013; Page & Duffy, 2018), health (Bhattacharya et al., 2012), and natural disaster events (Yang et al., 2019; Thomson et al., 2012). Although some work has been undertaken on determining credibility based on unsupervised approaches (Alrubaian et al., 2018), the related work on credibility assessment is comprised mainly of supervised approaches, which we now explore.

2.1. Tweet credibility

The majority of studies of credibility assessment on Twitter are comprised of supervised approaches, predominately decision trees, support vector machines, and Bayesian algorithms (Alrubaian et al., 2018). An extensive survey into the work of credibility on Twitter has been undertaken by Alrubaian et al. (2018), in which they looked at 112 papers on the subject of microblog credibility over the period 2006–2017. Alrubaian et al. (2018) cited one of the key challenges of credibility assessment is that there is a great deal of literature which has developed different credibility dimensions and definitions and that a unified definition of what constitutes credible information does not exist. This section will now explore the related work on supervised learning approaches for determining credibility, due to its popularity versus unsupervised approaches.

Castillo et al. (2011) were amongst the first to undertake research on the credibility of tweets, this work involved assessing the credibility of current news events during a two-month window. Their approach, which made use of Naïve Bayes, Logistic Regression, and Support Vector Machine, was able to correctly recognize 89% of topic appearances and their credibility classification achieved precision and recall scores in the range of 70–80%. Much of the work undertaken since has built upon the initial features proposed in this work. Morris et al. (2012) conducted a series of experiments which included identifying features which are highly relevant for assessing credibility. Their initial experiment found that there are several key features for assessing credibility, which include predominately user-based features such as the author's expertise of the particular topic being assessed (as judged by the author's profile description) and the user's reputation (verified account symbol). In a secondary experiment, they found that the topics of the messages influenced the perception of tweet credibility, with topics in the field of science receiving a higher rating, followed by politics and entertainment. Although the authors initially found that user images had no significant impact on tweet credibility, a follow-up experiment did establish that users who possess the default Twitter icon as their profile picture lowered credibility perception (Morris et al., 2012). Features which are derived from the author of the tweet have been studied intently within the literature, such features derived from the user have been criticised in recent works (Alrubaian et al., 2018)(Stringhini et al.,

Table 1
Related Supervised Research on Social Media Credibility.

Authors	Year	Num. of Microblog Posts Labelled	Annotation Strategy	Algorithm(s) Used	Num. of Features	Results
Hassan et al., (2018)	2018	5,802	Team of journalists – 2 labels (credible and not credible)	RF kNN SVM LR NB	32	79.6% precision (RF)
Ballouli et al., (2017)	2017	9,000	3 annotators 2 labels (credible and not credible)	RF NB SVM	48	66.8 – 76.1% precision (RF)
Krzysztof et al., (2015)	2015	1,206	2 annotators 4 labels (highly credible, highly non-credible, neutral, controversial)	SVM	12	84 – 89% precision (across the 4 classes)
F. Yang et al., (2012)	2012	5,155	2 annotators 2 labels (non-rumour and rumour)	RF	19	74.4 – 76.3% precision
C. Castillo et al., (2011)	2011	N/A – Tweets collected based on 2,500 topics	7 annotators (from crowdsourcing) 4 labels (almost certainly true, likely to be false, almost certainly true, I can't decide)	NB LR RF	30	89.1% precision (weighted average)

(RF – Random Forest, kNN – k-Nearest Neighbours, LR – Logistic Regression, NB – Naïve Bayes, SVM – Support Vector Machine)
Note: Results shown as based on the top-performing classifier.

2013), as features such as the number of followers a user has can be artificially inflated due to follower markets (De Micheli & Stroppa, 2013)(Cresci et al., 2015), indicating that feature could give a false indication of credibility.

Hassan et al. (2018) proposed a credibility detection model based on machine learning techniques in which an annotated dataset based on news events was annotated by a team of journalists. They proposed two features groups – content-based features (e.g. length of the tweet text) and source-based features (e.g. does the account have the default Twitter profile picture?) – in which classifiers were trained on features from each of these groups, and then trained on the combined feature groups. The results of this work showed that combining features from both groups led to performance gains versus using each of the feature sets independently. The authors, however, neglected to test that the performance between the two classifiers were statistically significant.

A summary of the previous work involving supervised approaches to assessing the credibility of microblog posts (Table 1) involves datasets annotated by multiple annotators. Bountouridis et al. (2019) studied the bias involved when annotating datasets in relation to credibility. They found that data biases are quite prevalent in credibility datasets. In particular, external, population, and enrichment biases are frequent and that datasets can never be neutral or unbiased. Like other subjective tasks, they are annotated by certain people, with a certain worldview, at a certain time, making certain methodological choices (Bountouridis et al., 2019). Studies often employ multiple annotators when a task is subjective, choosing to take the majority opinion of the annotators to reach a consensus (Sujoy Sikdar, Kang, O'donovan, Höllerer, & Adal, 2013; Castillo et al., 2011; Ballouli et al., 2017; Sikdar et al., 2014; Krzysztof et al., 2015), with some work removing observations in which a class cannot be agreed upon by a majority, or if annotators cannot decide upon any pre-determined label (Sujoy Sikdar, Kang, O'donovan, & Höllerer, 2013; Gupta & Kumaraguru, 2012).

Several other studies (Sikdar et al., 2014; Odonovan et al., 2012; Castillo et al., 2013) have focused on attempting to leverage the opinion of a large number of annotators through crowdsourcing platforms such as Amazon's Mechanical Turk[1] and Figure Eight[2] (formerly CrowdFlower). As annotators from crowdsourcing platforms tend not to know the message senders and likely do not have knowledge about the topic of the message, their ratings predominantly rely on whether the message text *looks* believable (Odonovan et al., 2012; Yang & Rim, 2014). Such platforms introduce other issues, in that such workers may not have previous exposure to the domain in which they are being asked to give a credibility rating to, and as a result, may not be invested in providing good-quality annotations (Hsueh et al., 2009). Alrubaian et al. (2018) also argue that depending on the wisdom of the crowd is not ideal, since a majority of participants may be devoid of related knowledge, particularly on certain topics which would naturally require prerequisite information (e.g. political events).

Although much of the supervised work on tweet credibility has been undertaken in an off-line (post-hoc) setting, some work has been undertaken on assessing the credibility of micro-blog posts in real-time as the tweets are published to Twitter. Gupta et al. (2014) developed a plug-in for the Google Chrome browser, which computes a credibility score for each tweet on a user's timeline, ranging from 1 (low) to 7 (high). This score was computed using a semi-supervised algorithm, trained on human labels obtained through crowdsourcing based on>45 features. The response time, usability, and effectiveness were evaluated on 5.4 million tweets. 63% of users of this plug-in either agreed with the automatically-generated score, as produced by the SVMRank algorithm or disagreed by 1 or 2 points.

2.2. Feature selection for credibility assessment

Much of the related work mentioned does not report on how informative each of the features are in their informative power to the classifiers, and simply just report the list of features and the overall metrics of the classifiers trained. Some of the features proposed previously in the literature could be irrelevant, resulting in poorer performance due to overfitting (Rani et al., 2015). Due to much of the related work not emphasising the importance of feature selection, this paper will attempt to address this shortcoming by emphasising the importance of effective feature selection methods. We will report on which features are the most deterministic, and which features are detrimental for assessing the financial credibility of microblogging tweets.

As the aforementioned previous works have explored, features are typically grouped up into different categories (e.g. tweet/content, user/author) and a credibility classification is assigned to a tweet, or to the author of the tweet. As a result of certain user features (e.g. number of followers a user has) being susceptible to artificial inflation, the methodology presented in this paper will assign a credibility to the tweet, and not make assumptions of the user and their background. With the related work on credibility assessment explored, the next section will present the methodology for assessing the credibility of financial stock tweets.

[1] https://www.mturk.com/
[2] https://www.figure-eight.com/

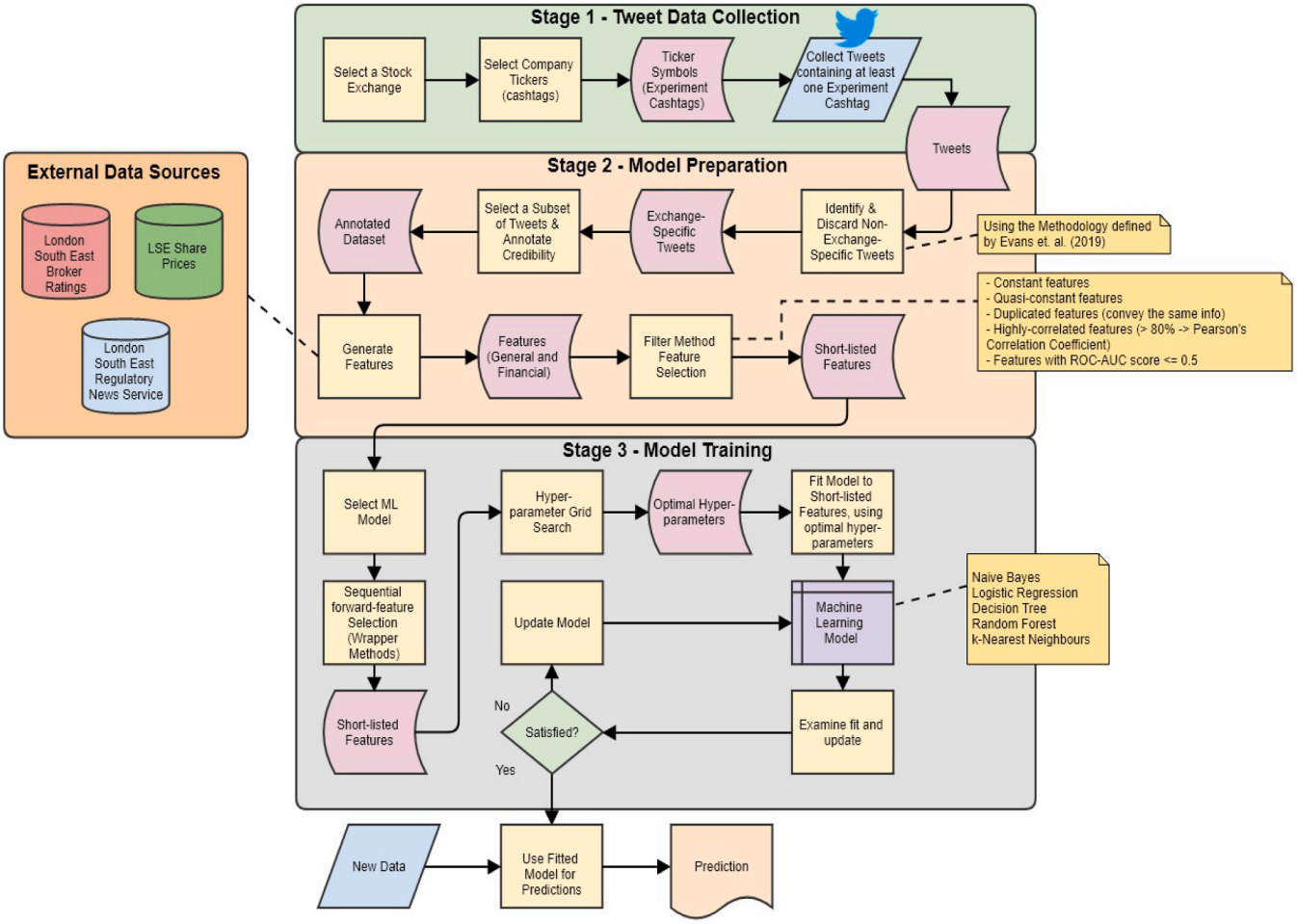

Fig. 1. Financial Credibility Assessment Methodology.

3. Methodology

Motivated by the success of supervised learning approaches in assessing the credibility of microblogging posts, we propose a methodology (Fig. 1) to assess the credibility of financial stock tweets (based on our definition of a stock tweet in Section 1). The methodology is comprised of three stages – the first stage of the methodology involves selecting a stock exchange in which to assess the credibility of financial stock tweets. With a stock exchange selected, a list of companies, and their associated ticker symbols can then be shortlisted in which to collect tweets. The second stage involves preparing the data for training machine learning classifiers by performing various feature selection techniques, explained in detail in Section 5. The final stage is the model training stage, in which models are trained on different feature groups with their respective performances being compared to ascertain if the proposed financial features result in more accurate machine learning models. This methodology will be validated by an experiment tailored for a specific stock exchange, explained further in Section 6. We now explain the motivation for each of these stages below.

3.1. Stage 1 – Data collection

The first step of the data collection stage is to select a stock exchange in which to collect stock tweets. Companies are often simultaneously listed on multiple exchanges worldwide (Gregoriou, 2015), meaning statements made about a specific exchange-listed company's share price may not be applicable to the entire company's operations. A shortlist of company ticker symbols can then be created to collect tweets for. Tweets can be collected through the official Twitter API (specific details discussed in Section 6.2). Once tweets have been collected for a given period for a shortlisted list of company ticker symbols (cashtags), tweets can be further analysed to determine if the tweet is associated with a stock-exchange listed company – the primary goal of the second stage of the methodology – discussed next.

3.2. Stage 2 – Model preparation

The second stage is primarily concerned with selecting and generating the features required to train the machine learning classifiers (Section 4) and to perform a quick screening of the features to identify those which are non-informative (e.g. due to being constant or highly-correlated with other features). Before any features can be generated,

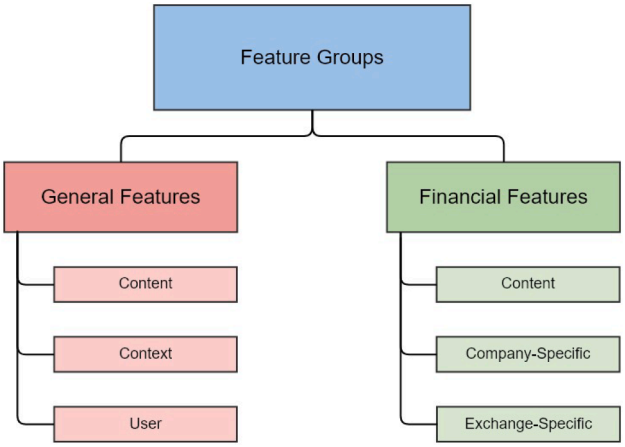

Fig. 2. Feature Subgroups.

however, it is important to note that identifying and collecting tweets for companies for a specific exchange is not always a straightforward task, as we will now discuss in the next subsection.

3.2.1. Identification of stock exchange-specific tweets

The primary issue of collecting financial tweets is that any user can create their own cashtag simply by prefixing any word with a dollar symbol ($). As cashtags mimic the company's ticker symbol, companies with identical symbols listed on different stock exchanges share the same cashtag (e.g. $TSCO refers to Tesco PLC on the London Stock Exchange, but also the Tractor Supply Company on the NASDAQ). This has been referred to as a cashtag collision within the literature, with previous work (Evans et al., 2019) adopting trained classifiers to resolve such collisions so that exchange-specific tweets can be identified, and non-stock-related market tweets can be discarded. We utilise the methodology of (Evans et al., 2019) to ensure the collection of exchange-specific tweets and is considered a data cleaning step. Once a suitable subsample of tweets has been obtained after discarding tweets not relating to the pre-chosen exchange, features can then be generated for each of the observations.

3.2.2. Dataset annotation

As supervised machine learning models are to be trained, a corpus of tweets must be annotated based on a pre-defined labelled system. As discussed in the related work on supervised learning approaching for credibility assessment (Section 2.1), this is sometimes approached as a binary classification problem (i.e. the tweet is either credible or not credible), with some work opting for more granularity of labels by incorporating labels to indicate the tweet does not have enough information to provide a label in either direction. Section 6.3 includes a detailed overview of the annotation process undertaken for the experiment within this paper.

3.2.3. Feature engineering and selection

After an annotated dataset has been obtained, the features can be analysed through appropriate filter-based feature selection techniques in an attempt to reduce the feature space, which may result in more robust machine learning models (Rong et al., 2019). Such filter methods include identifying constant or quasi-constant features, duplicated features which convey the same information, and features which are highly correlated with one another (Bommert et al., 2020). Section 5 provides a detailed overview of each of the feature methods in this work.

3.3. Stage 3 – Model training

The final stage of the methodology involves further feature selection techniques (discussed in Section 5) through repeated training of classifiers to discern optimal feature sets by adopting techniques such as wrapper methods. Once an optimal feature subset has been identified, the methodology proposes performing a hyperparameter grid search to further improve the performance of the various classifiers. Although the methodology proposes training traditional supervised classifiers, this list is not exhaustive and can be adapted to include other supervised approaches. The next section introduces the proposed general and financial features to train the machine learning models.

4. Proposed features

Many of the general features (GF) we propose have been used in previous work on the assessment of tweet credibility (Alrubaian et al., 2018). The full list of proposed features (both general and financial), along with a description of each feature can be found in Appendix A. We concede that not every feature proposed will offer an equal amount of informative power to a classification model, and as a result, we do not attempt to justify each of the features in turn, but instead remove the feature(s) if they are found to be of no informative value to the classifiers. The general and financial feature groups, including their associated sub-groups, are provided in Fig. 2.

4.1. General features (GF)

The GF group is divided into three sub-groups – content, context, and user. Content features are derived from the viewable content of the tweet. Context features are concerned with information relating to *how* the tweet was created, including the date and time and source of the tweet. User features are concerned with the author of the tweet. Each of these sub-groups will now be discussed further.

4.1.1. Content

Content-derived features are features directly accessible from the tweet text or can be engineered from the tweet text. The features proposed in this group include the count of different keyword groups (e.g. noun, verb) and details of the URLs found within the tweet. Many of the features within this group assists in the second dimension of financial tweet credibility – how credible is the information within the tweet?

4.1.2. Context

Features within the context sub-group include when the tweet was published to Twitter, in addition to extracting the number of live URLs from the tweet. We argue that simply the presence of a URL should not be seen as a sign of credibility, as it could be the case that the URL is not active in the sense it redirects to a web server. The count of live URLs within the tweet (F27 - Table A1) involves visiting each of the URLs in the tweet to establish if the URL is still live. We define a live URL as any URL which returns a successful response code (200). The number of popular URLs within the tweet, as determined by the domain popularity ranking website, moz[3].

Tweets can be published to Twitter in a variety of ways – these can typically be grouped into manual or automatic. Manual publishing methods involve the user manually publishing a tweet to Twitter, whereas automatic tweets are published based on rules and triggers (Castillo et al., 2019), such as a specific time of the day. Many providers exist for the automatic publishing of content to Twitter (Saguna et al., 2012), such as TweetDeck, Hootsuite, IFTTT. The Tweet Source feature is encoded based on which approach was used to publish the tweet, as described in Table A1.

[3] https://moz.com/top500

Table 2
Financial Keyword Groups (as defined by (Loughran et al., 2011)).

Keyword Group	Group Description	Total Number of Keywords in Group	Example Keywords
Positive	Positive in a financial setting	354	booming, delighted, encouraged, excited, lucrative, meritorious, strong, winner
Negative	Negative in a financial setting	2355	abnormal, aggravated, bankruptcy, bribe, challenging, defamation, disaster
Uncertainty	Indicates uncertainty	297	anomalous, could, fluctuation, probable, random
Litigious	Indicates litigious action	904	claimholder, testify, whistleblower, voided, ruling, perjury
Constraining	Words indicating constraints, (debt, legal, employee, and environmental)	194	compel, depend, indebted, mandate, pledge, prevent, refrain, strict, unavailable

4.1.3. User

Used extensively within the literature for assessing credibility (Alrubaian et al., 2018), user features are derived or engineered from the user authoring the tweet. This feature group assists with the third dimension of financial tweet credibility – how credible is the author of the tweet? The proposed user features to be used in the methodology involve how long a user has been active on Twitter at the time a tweet was published (F31) and details on their network demographic (follower/following count). As discussed in Section 2.1, previous work (Morris et al., 2012) found that users possessing the default profile image were perceived as less credible.

4.2. Financial features (FF)

We now present an overview of the FF proposed for assessing the financial credibility of stock tweets. FF are further divided into three groups: content, company-specific, and exchange-specific. As discussed in Section 1, the financial features proposed (Table A2) are novel in that they have yet to be proposed in the literature. We hypothesise that the inclusion of such features will contribute to improved performance (over classifiers trained on general or financial features alone) when combined with the GF proposed in Section 4.1. Many of these features are dependent on external sources relating to the company corresponding to the tweet's cashtag (such as the range of the share price for that day), including the exchange in which the company is listed on (e.g. was the stock exchange open when the tweet was published). These FF will now be discussed further, beginning with the features which can be derived from the content of the tweet.

4.2.1. Content

Although many sentiment keyword lists exist for the purpose of assessing the sentiment of text, certain terms may be perceived differently in a financial context. If word lists associate the terms *mine*, *drug*, and *death* as negative, as some widely used lists do (Loughran & Mcdonald, 2016), then industries such as mining and healthcare will likely be found to be pessimistic. Loughran et al. (2011) have curated keyword lists which include positive, negative, and uncertainty keywords in the context of financial communication. This keyword list (summarised in Table 2) contains over 4,000 keywords and was obtained using standard financial texts. Each of the keyword categories is transformed into its own respective feature (see F45-F49 in Table A2). There are other lexicons available which have been adapted for microblogging texts (Oliveira et al., 2016; Houlihan & Creamer, 2019), which could be also be effective to this end. However, we elect to use the lexicon constructed by Loughran et al. (2011) due to it being well-established within the literature.

4.2.2. Company-specific

Stock prices for exchange-listed companies are provided in open, high, low, and close (OHLC) variants. These can either be specific to a certain time window, such as every minute, or to a period such as a day. We propose two features which are engineered from these price variants – the range of the high and low price for the day (F50) the tweet was made, and the range of the close and open price (F51).

4.2.3. Exchange-specific

Several of the FF proposed differ slightly depending on the stock exchange in question. The number of credible financial URLs in the tweet (F54) requires curating a list of URLs which are renowned as being a credible source of information. Several other features proposed (F55-F56) involve establishing if the tweet was made when the stock exchange was open or closed – different stock exchanges have differing opening hours, with some closing during lunch. The next section will discuss the feature selection techniques to be adopted by the methodology.

5. Feature selection

Naturally, not each of the features proposed in Appendix A will provide informative power to all machine learning classifiers. It is, therefore, appropriate to perform appropriate feature selection techniques to assess how informative each of these features are. Sometimes, a large number of features may lead to models which overfit, leading them to reach false conclusions and negatively impact their performance (Arauzo-Azofra et al., 2011). Other benefits of feature selection include improving interpretability and lowering the cost of data acquisition and handling, thus improving the quality of such models. It is also prudent to note that not every classifier will benefit from performing feature selection. Decision trees, for instance, have a feature selection mechanism embedded within them where the feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can then be calculated by the number of samples that reach that node, divided by the total number of samples – with higher values indicating the importance of the feature (Ronaghan, 2018). Random Forest classifiers also naturally share this mechanism of feature selection. Other machine learning models often employ some kind of regularization that punish model complexity and drive the learning process towards robust models by decreasing the less impactful feature to zero and then dropping them (e.g. Logistic Regression with L1-regularization) (Coelho & Richert, 2015).

5.1. Filter methods

Often used as a data pre-processing step, filter methods are based on statistical tests which are performed prior to training machine learning models. The goal of filter methods is to identify features which will not offer much, or any, informative power to a machine learning model. Such methods are aimed at finding features which are highly correlated or features which convey the exact same information (duplicated). Filter

Table 3
Annotated Tweet Breakdown.

Label	Meaning	Count of Annotated Tweets	Count when Merged
0	Strong Not Credible	814	2134
1	Not Credible	1320	
2	Ambiguous/Not enough Info	693	693
3	Fairly Credible	1020	2173
4	Very Credible	1153	

methods can be easily scaled to high-dimensional datasets, are computationally fast and simple to perform, and are independent of the classification algorithms to which they aim to improve (Tsai & Chen, 2019). Different filter methods exist and perform differently depending on the dimensionality and types of datasets. A detailed overview of the different types of filter methods available for high-dimensional classification data can be found in (Bommert et al., 2020).

5.2. Wrapper methods

Wrapper methods are also frequently used in the machine learning process as part of the feature selection stage. This technique aims to find the best subset of features according to a specific search strategy (Dorado et al., 2019). Popular search strategies include sequential forward feature selection, sequential backward feature selection, and recursive feature elimination. As such wrapper methods are designed to meet the same objective – to reduce the feature space – any of these techniques can be adopted to meet this end.

6. Experimental design

In order to validate the credibility methodology (Section 3), an experiment has been designed using tweets relating to companies listed on the London Stock Exchange (LSE). This experiment will follow the suggested steps and features proposed in the methodology for assessing the financial credibility of tweets (Section 4.2).

6.1. Company selection

Before collection of the tweets can commence, the ticker symbols of companies need to be determined. The LSE is divided into two secondary markets; the Main Market (MM), and the Alternative Investment Market (AIM). Each exchange-listed company belongs to a pre-defined industry: basic materials, consumer goods, consumer services, financials, health care, industrials, oil & gas, technology, telecommunications, and utilities. We have selected 200 companies (100 MM, 100 AIM) which have been listed on the LSE for at least two years (to give an optimal chance that tweets can be collected for that cashtag, and therefore the company), these companies are referred to as the experiment companies in the rest of this paper and can be viewed in Appendix B.

6.2. Data collection

Twitter provides several ways to collect tweets. The first is from Twitter's Search API, which allows the collection of tweets from up to a week in the past for free. Another way is to use the Twitter Streaming API (Nguyen et al., 2015), allowing the real-time collection of tweets. We have collected tweets containing at least one occurrence of a cashtag of an experiment company. In total, 208,209 tweets were collected over a one-year period (15/11/19 – 15/11/20). Several of the features proposed in Appendix A require that the data be retrieved from external APIs. The daily share prices for each experiment company has been collected from AlphaVantage for the date. Broker ratings and dates in which Regulatory News Service notices were given have been web scraped from London South East, a website which serves as an aggregator for financial news for the LSE for the dates covering the data collection period.

Table 4
Inter-Item Correlation Matrix & CA Scores for binary-labelled tweets. CA = 0.591 (Sample size = 10).

	MA	A1	A2	A3	CA if item deleted
MA	1.000	−0.200	0.816	0.816	0.148
A1	−0.200	1.000	0.000	−0.408	0.895
A2	0.816	0.000	1.000	0.583	0.179
A3	0.816	−0.408	0.583	1.000	0.433

Table 5
Inter-Item Correlation Matrix & CA Scores for five-class labelled tweets. CA = 0.699 (Sample size = 10).

	MA	A1	A2	A3	CA if item deleted
MA	1.000	−0.061	0.722	0.827	0.443
A1	−0.061	1.000	0.210	−0.063	0.866
A2	0.722	0.210	1.000	0.578	0.538
A3	0.827	−0.063	0.578	1.000	0.518

Table 6
Inter-Item Correlation Matrix & CA Scores for three-class labelled tweets. CA = 0.686 (Sample size = 30).

	MA	A1	A2	A3	CA if item deleted
MA	1.000	0.715	0.752	0.173	0.449
A1	0.715	1.000	0.600	0.052	0.547
A2	0.752	0.600	1.000	0.055	0.537
A3	0.173	0.052	0.055	1.000	0.866

6.3. Tweet annotation

After tweets containing at least one occurrence of an experiment company's cashtag, a subsample of 5,000 tweets were selected. We began by attempting to retrieve 25 tweets for each experiment company cashtag, this resulted in 3,874 tweets – tweets were then randomly selected to reach a total of 5,000 tweets.

As discussed in Section 2.1, subjective tasks such as annotating levels of credibility can vary greatly depending on the annotators' perceptions. Any dataset annotated by an individual which is then used to train a classifier will result in the classifier learning the idiosyncrasies of that particular annotator (Reidsma and op den Akker, 2008). To alleviate such concerns, we began by having a single annotator (referred herein as the main annotator – MA) provide labels for each tweet based on a five-label system (Table 3). We then take a subsample (10) of these tweets and get the opinion of three other annotators who have had previous experience with Twitter datasets, to ascertain the inter-item correlation between the annotations. To assess the inter-item correlation, we compute the Cronbach's Alpha (CA) (Eq. (1)) of the four different annotations for each of the tweets.

$$\alpha = \frac{N\bar{c}}{\bar{v} + (N-1)\bar{c}} \tag{1}$$

where N is the number of items, $\bar{c}$ is the average inter-item covariance among the items and $\bar{v}$ is the average variance. A Cronbach score of >0.7 infers a high agreement between the annotators (Landis & Koch, 1977). The CA for the binary labelled tweets (Table 4) – 0.591 – shows that the four annotators were unable to reach a consensus as to what constitutes a credible or not credible tweet. The CA for the five-label system (Table 5) – 0.699 – shows that annotators were able to find a more consistent agreement, although it did not meet the threshold of constituting a high agreement. A further experiment involving a three-label scale (not credible, ambiguous, and credible), with a larger sample

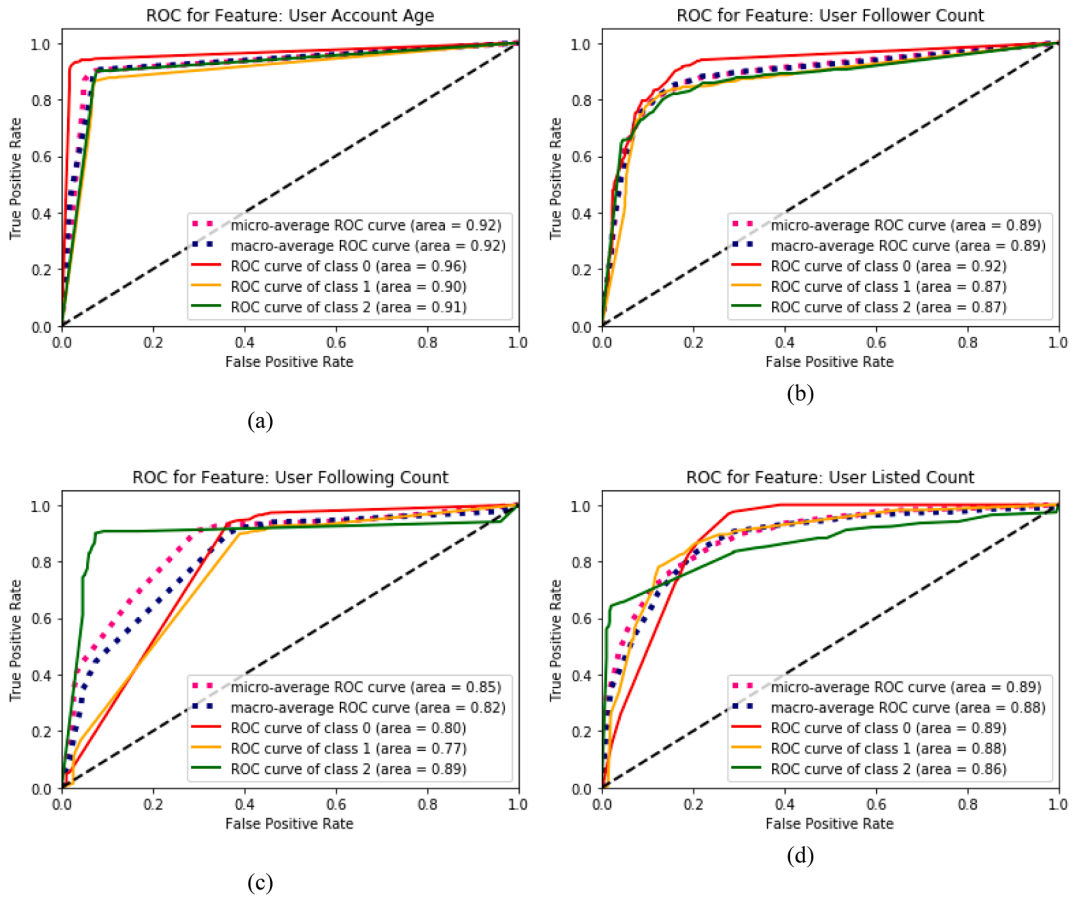

Fig. 3. Top Four Features based on Macro-AUC.

Table 7
Non-Informative Features.

Feature Selection Technique	Description	Features Identified
Constant features	Features which are constant among all observations	Tweet contains pos emoticons Tweet contains neg emoticons
Quasi-constant features	Features which are constant amongst almost all observations.	Tweet contains multiple question marks Tweet contains exclamation mark Tweet contains exclamation Count of second-person pronouns User is verified Tweet is a quote tweet Contains media Interjection word count Constraining keyword count
Duplicated features	Features which convey the same information	*None*
Highly-correlated features	Features with a Pearson's correlation coefficient of > 0.8	User has non-fictional location Is RT Tweet Length (Words) Username word count
Univariate ROC-AUC score	Features which have a ROC-AUC score close to random chance	Financial CTs Technology CTs Telecommunication CTs

size of 30 tweets, was then performed to assess the annotators' agreement on such a scale. In each of these experiments, it is clear that if the CA is computed with the MA removed, it results in the greatest decrease in the CA score – indicating the majority of the annotators' opinions are mostly aligned to that of the MA. Although none of these experiments results in a CA of > 0.7, we seek to find a consensus with the majority annotators, provided that the MA is not in the minority. The highest CA score (from the majority $-$ 3) comes from the binary-labelled system, in which if A1 is removed, the CA becomes 0.895, indicating the MA, A2 and A3 have reached a consensus on annotating credibility. A binary label approach, however, does not offer the granularity which is often achieved versus a multiclass approach. As the five-class system has a significant class imbalance when taking into consideration the individual classes (814 strong not credible vs 1320 not credible tweets), We have elected to adopt the three-class approach which combines the two not-credible classes and the two credible classes, and to ensure that ambiguous tweets can be taken into consideration (Table 6).

6.4. *Assessing feature importance*

As discussed in Section 5, assessing the informative power of each of the features in isolation can help remove features which will not positively affect the performance of the machine learning classifiers. To this end, for each feature, a Decision Tree (DT) classifier has been trained to assess the importance of the feature when predicting each of the classes. The metric used to calculate the importance of each feature is the probability returned from the DT. We then calculate the total area under the curve (AUC) for the feature. Naturally, the AUC can only be computed for a binary classification problem. In order to calculate the

Table 8
Classifier Results.

	General Features						Financial Features						General + Financial Features					
Classifier	Features (/34)	Acc	Pre	Rec	F1	AUC	Features (/21)	Acc	Pre	Rec	F1	AUC	Features (/55)	Acc	Pre	Rec	F1	AUC
NB	4	85.5	84.8	85.5	85.0	89.1	12	61.0	63.9	60.3	59.7	70.4	6 (2FF)	85.6	84.9	85.6	85.1	91.4
LR	21	88.0	84.6	86.0	85.3	90.5	9	55.9	40.8	50.7	43.0	64.0	27 (9FF)	87.6	87.1	86.8	86.9	92.0
DT	18	90.1	90.6	90.4	90.5	92.6	10	54.2	55.1	49.6	43.0	63.1	11 (3FF)	89.7	90.1	90.0	90.0	93.1
RF	20	92.7	93.1	92.6	92.9	93.8	11	61.9	63.1	60.9	60.4	70.9	37 (12FF)	**93.5**	**94.3**	**93.2**	**93.7**	**94.3**
kNN	7	91.4	92.3	91.1	91.6	93.2	7	61.5	64.0	61.3	60.8	71.1	9 (2FF)	92.7	93.6	92.5	92.9	93.6

Note: Scores presented are the macro average percentage (%).

AUC for a multi-class problem, the DT classifier, which is capable of producing an output $y = \{0, 1, 2\}$, is converted into three binary classifiers through a One-Vs-Rest approach (Ambusaidi et al., 2016). Each of the AUC scores for the three binary classifiers, for each feature, can then be calculated to ascertain the feature's predictive power for each class. The AUC score can be computed in different ways for a multiclass classifier: the macro average computes the metric for each class independently before taking the average, whereas the micro average is the traditional mean for all samples (Aghdam et al., 2009). Macro-averaging treats all classes equally, whereas micro-averaging favours majority classes. We elect to judge the informative power of the feature based on its AUC macro average, due to ambiguous tweets being relatively more uncommon than credible and not credible tweets. Four of the features (Fig. 3) exhibit a macro AUC score of > 0.8, indicating they will likely offer a great degree of informative power when used to train machine learning classifiers. These four features are all contained within the general group and are attributed to the user of the tweet, and is consistent with previous work (Yang et al., 2012) which found that user attributes to be incredibly predictive of credibility.

The filter methods outlined in the methodology (Fig. 1), have been applied to the annotated dataset (5,000 tweets). Based on these five different filter method feature selection techniques, 18 features (Table 7) have been identified to provide no meaningful informative power based on the probability returned from the DT.

With the informative and non-informative features indentified, machine learning classifiers can now be trained on an optimal feature set. The 18 non-informative features identified have been dropped due to the reasons outlined in Table 7.

7. Experimental results & discussion

We now present the results (Table 8) obtained from the experiment based on all of the features after the non-informative features are removed (34 GF, 21 FF), and illustrate that some models' performance suffers if feature selection techniques are not taken into consideration. We have trained classifiers which have demonstrated previous success in assessing the credibility of microblog messages (Naïve Bayes, k-Nearest Neighbours, Decision Trees, Logistic Regression, and Random Forest) (Alrubaian et al., 2018). All of the results obtained are a result of 10-fold cross-validation using an 80/20 train/test split and implemented using the scikit-learn library within Python. Each of the classification models underwent a grid search to find optimal hyperparameters. Three sets of classifiers have been trained; (1) trained on the GF, (2) trained on the FF, and (3) trained on both sets of features.

As indicated by the results of the sequential feature selection (Fig. 4), the kNN and NB classifiers suffer clear decreases in their performance when more features are added to the feature space due to the well-documented phenomenon of the curse of dimensionality (Parmezan et al., 2017). DT, RF, and LR, also suffer minor decreases, although, due to the nature of these three algorithms, they are less impacted. Based on the AUC, the RF classifier is the top-performing classifier when trained on the GF and FF sets respectively. Clearly, classifiers trained solely on the FF pale in performance when compared to classifiers trained on the other feature sets. Regarding RQ1, GF by themselves are extremely informative for assessing the credibility classification of tweets. When combined with FF (RQ2), performance gains are evident in all of the classifiers trained on the combined feature sets. The importance of feature selection is particularly prevalent for the kNN classifier, which reaches its zenith at 9 features and almost outperforms the RF when both are compared at such a feature space size. In terms of which FFs were seen to be informative, the RF trained on the combined features utilised 12 financial features, which included; F46, F55, F56, F58, and 8xF59+). In respect to the five classifiers trained on the combined features, the most popular FFs utilised by the classifiers were the count of cashtags in the tweet (F58), and the count of technology and healthcare cashtags within the tweet (2xF59+).

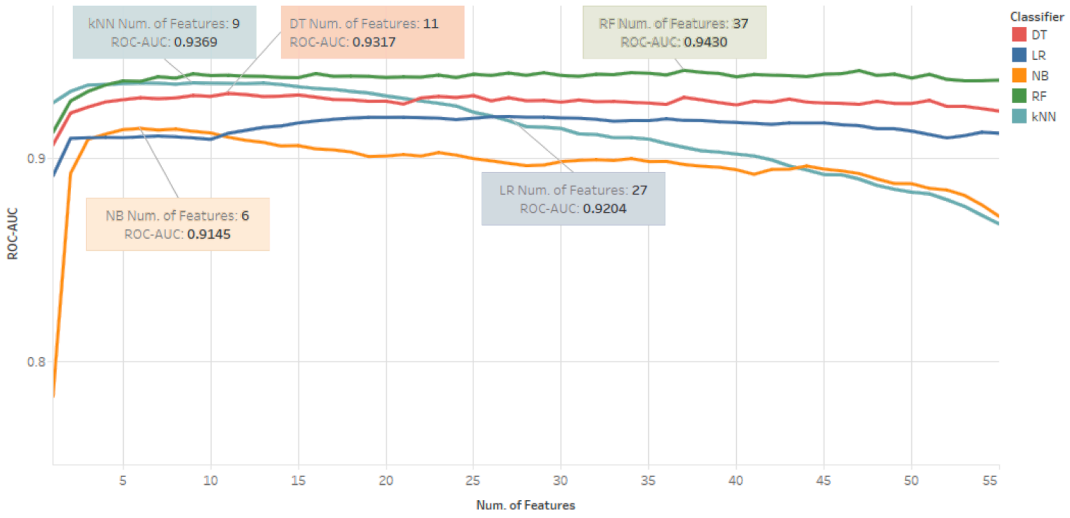

Fig. 4. Sequential Forward Feature Selection Results (Combined features).

As evident from the initial experiment results, RF appears to be the best performing classifier when the feature sets are combined. We now test if the differences between the predictions of the RF trained on GF versus the RF trained on the combined features are statistically significant by conducting the Stuart-Maxwell test. The Stuart-Maxwell test is an extension to the McNemar test, used to assess marginal homogeneity in independent matched-pair data, where responses are allowed more than two response categories (Yang et al., 2011). The p-value of the Stuart-Maxwell test on the predictions of both the RF trained on GF and the RF trained on the combined features is 0.0031, indicating the difference between the two classifiers are statistically significant.

8. Conclusion

This paper has presented a methodology for assessing the credibility of financial stock tweets. Two groups of features were proposed, GF widely used within the literature and a domain-specific group specific to financial stock tweets. Before the training of classifiers, feature selection techniques were used to identify non-informative features. Based on the two groups of features (general and financial), three sets of classifiers were trained, with the first two groups being the set of general and FF respectively, and the third being the combination of the two. Performance gains were noted in the machine learning classifiers, with some classifiers (NB and kNN) suffering when their respective feature spaces grew, undoubtedly due to the curse of dimensionality. Although the RF classifiers were certainly the best performing classifiers in respect to the AUC, it is important to note that the kNN classifier trained on the combined feature set was also a formidable classifier due to its comparative performance with the RF classifiers without having to take into account as many features (9 features compared to 37 for RF). The number of dependent features for the RF classifier presents some limitations for deploying a model dependent on a larger number of features, some of which are more computationally to obtain than others. The count of live URLs within the tweet (F27) requires querying each URL in the tweet, which can be computationally expensive to generate the feature if a tweet contains multiple URLs. Establishing the computational cost of features such as the count of live URLs in a tweet and to assess their suitability in a real-time credibility model is an interesting avenue for future work. There are other features which could be engineered by querying external APIs such as historical stock market values and ascertaining if the tweet contains credible information regarding stock movements of the cashtags contained in the tweet. This would be most beneficial if attempting to classify user credibility – does a user often tweet information about stock-listed companies which turned out to be true? Adopting a lexicon which has been constructed based on financial microblog texts, such as the one constructed by (Oliveira et al., 2016) could yield improved results when assessing tweet credibility, this is an avenue for future work.

As discussed in section 3.3, the list of supervised classifiers in this work is not exhaustive, Support Vector Machines (SVM) were included in the list of classifiers to be trained, but performing hyperparameter grid searches were extremely computationally expensive and were abandoned due to the unsuitability of comparing the SVM classifier with no hyperparameter tuning to that of models which had undergone extensive hyperparameter tuning. Future work in this regard would include the SVM to assess its predictive power in classifying the credibility of financial stock tweets, with neural network architectures also being considered. The credibility methodology presented in this paper will be utilised in the future by a smart data ecosystem, with the intent of monitoring and detecting financial market irregularities.

CRediT authorship contribution statement

Lewis Evans: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Majdi Owda:** Conceptualization, Methodology, Validation, Writing - review & editing, Supervision, Project administration. **Keeley Crockett:** Conceptualization, Methodology, Validation, Writing - review & editing, Supervision, Project administration. **Ana Fernandez Vilas:** Conceptualization, Methodology, Validation, Writing - review & editing, Supervision, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A

Table A1

Table A1

Feature Sub-Group	Feature Num.	Feature	Notes
Content	1	Tweet Length (Chars)	Length of the tweet in characters (including spaces)
	2	Tweet Length (Words)	Length of the tweet in words
	3	Tweet Contains Question Mark (QM)	Does the tweet contain a question mark
	4	Tweet Contains Multiple QMs	Does the tweet contain multiple question marks
	5	Tweet Contains Exclamation Mark (EM)	Does the tweet contain an exclamation mark
	6	Tweet Contains Multiple EMs	Does the tweet contain multiple exclamation marks
	7	Tweet Contains First Person Pronouns	e.g. I, we, us, me, my, mine, our, ours
	8	Tweet Contains Second Person Pronouns	e.g. you, your, yours
	9	Tweet Contains Third Person Pronouns	e.g. he, she, her, him, it, they, them, theirs
	10	Tweet Contains Positive Emoticons	e.g. :), :-)
	11	Tweet Contains Negative Emoticons	e.g. :(, :-(
	12	Tweet Contains User Mention	Does the tweet contain an @ user mention
	13	Tweet Hashtag Count	The count of word prefixed with a hashtag (#) as determined by the tweet JSON object
	14	Is Retweet (RT)	Contains RT at the start of the tweet text
	15	URL Count	The count of URLs within the tweet
	16	Per cent Uppercase	The percentage of the tweet which is in UPPERCASE
	17	Is Quote Tweet	If the tweet is quoting (e.g. replying) to another tweet
	18	Contains Media	Contains an image, video or gif
	19	Present Verb Count	Count of verbs in present tense within the tweet text
	20	Past Verb Count	Count of verbs in past tense within the tweet text
	21	Adjective Count	Count of adjectives within the tweet text
	22	Interjection Count	Count of interjections within the tweet text
	23	Noun Count	Count of nouns within the tweet text
	24	Adverb Count	Count of adverbs within the tweet text
	25	Proper Noun Count	Count of proper nouns within the tweet text
	26	Numerical Cardinal Count	Count of numerical cardinal values within the tweet text
Context	27	Live URL Count	The count of URLs in the tweet which resulted in a successful web response (200)
	28	Tweeted on Weekday	If the tweet was tweeted on a weekday
	29	Top 500 URL Count	As defined by https://moz.com/top500
	30	Tweet Source	0 – Official Twitter Web Client1 – Twitter for Android2 – Twitter for iPhone3 – Automated Tool (e.g. Zapier, IFTTT, Hootsuite, TweetDeck)4 – Other
User	31	User Account Age (at time of tweet)	The number of days an account has been active on the Twitter platform from when the tweet was published to Twitter
	32	User has URL on Profile	Does the user have a URL on their profile?
	33	User has Default Profile Pic	Is the user using the default profile image provided by Twitter upon registering their account
	34	User has set a Location	Has the user set a location on their profile?
	35	User Verified	Is the user a verified user (blue tick verification seal)?
	36	User Num of Tweets	The number of tweets the user has made (at the time the tweet was collected)
	37	User Follower Count	The number of followers the user's account has
	38	User Following Count	The number of accounts the user is following
	39	User Listed Count	How many lists is the user account's listed on?
	40	User has Desc	Does the user have a description on their profile page?
	41	User Description Length	The length of the user description, 0 if none
	42	User has Real Location	Does the user have a factual location?
	43	Username Length	Length of the user's username
	44	Username Words	The number of words comprising the user name

Table A2

Table A2
Financial Feature List.

Feature Sub-Group	Feature Num.	Feature	Notes
Content	45	Count of positive financial keywords	As defined by research by (Loughran et al., 2011).
	46	Count of negative financial keywords	
	47	Count of uncertainty financial keywords	
	48	Count of litigious financial keywords	
	49	Count of constraining financial keywords	
Company-Specific Features	50	Close – Open Price (range) on day	Provided by the AlphaVantage API
	51	High – Low Price (range) on day	
	52	RNS published on day	Was a Regulatory News Service (RNS) statement issued for the company corresponding to the first experiment cashtag encountered on the day the tweet was made?
	53	Broker Rating issued on day	Was a Broker rating issued for the company corresponding to the first experiment cashtag encountered on the day the tweet was made?
Exchange-Specific Features	54	Credible Fin URLs in Tweet	A list of URLs found to be credible investment or news websites, hand-curated by an expert based on all the URLs found occurring in at least 1% of the overall tweets collected.
	55	Tweeted Before Market Open	These features differ depending on the stock exchange.
	56	Tweeted During Market Open	
	57	Tweeted After Market Closed	
	58	Count Cashtags (CTs)	
	59+	Count of each industry Cashtags	

Appendix B

Table B1

Table B1
Experiment Companies (AIM-listed).

Company Ticker	Company Name	Company Industry
GGP	Greatland Gold Plc	Basic Materials
VRS	Versarien Plc	Basic Materials
KDNC	Cadence Minerals Plc	Basic Materials
BIOM	Biome Technologies Plc	Basic Materials
CRPR	Cropper (James) Plc	Basic Materials
PREM	Premier African Minerals Limited	Basic Materials
AAU	Ariana Resources Plc	Basic Materials
RRR	Red Rock Resources Plc	Basic Materials
HRN	Hornby Plc	Consumer Goods
MUL	Mulberry Group Plc	Consumer Goods
WYN	Wynnstay Group Plc	Consumer Goods
FEVR	Fevertree Drinks Plc	Consumer Goods
TUNE	Focusrite Plc	Consumer Goods
LWRF	Lightwaverf Plc	Consumer Goods
FDEV	Frontier Developments Plc	Consumer Goods
G4M	Gear4music (Holdings) Plc	Consumer Goods
HOTC	Hotel Chocolat Group Plc	Consumer Goods
SIS	Science In Sport Plc	Consumer Goods
TEF	Telford Homes Plc	Consumer Goods
ZAM	Zambeef Products Plc	Consumer Goods
ASC	Asos Plc	Consumer Services
EMAN	Everyman Media Group Plc	Consumer Services
JOUL	Joules Group Plc	Consumer Services
BOO	Boohoo.Com Plc	Consumer Services
KOOV	Koovs Plc	Consumer Services
YOU	Yougov Plc	Consumer Services
APGN	Applegreen Plc	Consumer Services
CCP	Celtic Plc	Consumer Services
CRAW	Crawshaw Group Plc	Consumer Services
FJET	Fastjet Plc	Consumer Services
SHOE	Shoe Zone Plc	Consumer Services
TMO	Time Out Group Plc	Consumer Services
UCG	United Carpets Group Plc	Consumer Services

(*continued on next page*)

Table B1 (*continued*)

Company Ticker	Company Name	Company Industry
HUNT	Hunters Property Plc	Financials
MTR	Metal Tiger Plc	Financials
CRC	Circle Property Plc	Financials
BLV	Belvoir Lettings Plc	Financials
TUNG	Tungsten Corporation Plc	Financials
PURP	Purplebricks Group Plc	Financials
ARGO	Argo Group Limited	Financials
MTW	Mattioli Woods Plc	Financials
TPFG	Property Franchise Group Plc (The)	Financials
PGH	Personal Group Holdings Plc	Financials
MAB1	Mortgage Advice Bureau (Holdings) Plc	Financials
ABC	Abcam Plc	Health Care
COG	Cambridge Cognition Holdings Plc	Health Care
AMYT	Amryt Pharma Plc	Health Care
CLIN	Clinigen Group Plc	Health Care
HZD	Horizon Discovery Group Plc	Health Care
AGL	Angle Plc	Health Care
AVCT	Avacta Group Plc	Health Care
KMK	Kromek Group Plc	Health Care
REDX	Redx Pharma Plc	Health Care
SUN	Surgical Innovations Group Plc	Health Care
SAR	Sareum Holdings Plc	Health Care
FLOW	Flowgroup Plc	Industrials
INSE	Inspired Energy Plc	Industrials
NAK	Nakama Group Plc	Industrials
DX.	Dx (Group) Plc	Industrials
WYG	Wyg Plc	Industrials
MRS	Management Resource Solutions Plc	Industrials
ASY	Andrews Sykes Group Plc	Industrials
BEG	Begbies Traynor Group Plc	Industrials
CTG	Christie Group Plc	Industrials
GTLY	Gateley (Holdings) Plc	Industrials
UTW	Utilitywise Plc	Industrials
88E	88 Energy Limited	Oil & Gas
GBP	Global Petroleum Limited	Oil & Gas
ITM	Itm Power Plc	Oil & Gas
CLON	Clontarf Energy Plc	Oil & Gas
NAUT	Nautilus Marine Services Plc	Oil & Gas
SOU	Sound Energy Plc	Oil & Gas
ANGS	Angus Energy Plc	Oil & Gas
HUR	Hurricane Energy Plc	Oil & Gas
NUOG	Nu-Oil And Gas Plc	Oil & Gas
TLOU	Tlou Energy Limited	Oil & Gas
SLE	San Leon Energy Plc	Oil & Gas
EYE	Eagle Eye Solutions Group Plc	Technology
ING	Ingenta Plc	Technology
TRB	Tribal Group Plc	Technology
BGO	Bango Plc	Technology
WAND	Wandisco Plc	Technology
PRSM	Blue Prism Group Plc	Technology
ALB	Albert Technologies Ltd	Technology
AMO	Amino Technologies Plc	Technology
BBSN	Brave Bison Group Plc	Technology
ESG	Eservglobal Limited	Technology
FBT	Forbidden Technologies Plc	Technology
IOM	Iomart Group Plc	Technology
RDT	Rosslyn Data Technologies Plc	Technology
TCM	Telit Communications Plc	Technology
ZOO	Zoo Digital Group Plc	Technology
AVN	Avanti Communications Group Plc	Telecommunications
MANX	Manx Telecom Plc	Telecommunications
GAMA	Gamma Communications Plc	Telecommunications
MOS	Mobile Streams Plc	Telecommunications
TPOP	People's Operator Plc (The)	Telecommunications
GOOD	Good Energy Group Plc	Utilities
YU.	Yu Group Plc	Utilities
ACP	Armadale Capital Plc	Utilities

Table B2

Table B2
Experiment Companies (MM-listed).

Company Ticker	Company Name	Company Industry
ACA	Acacia Mining Plc	Basic Materials
BFA	BASF Se	Basic Materials
BLT	BHP Billiton Plc	Basic Materials
PDL	Petra Diamonds Limited	Basic Materials
RIO	Rio Tinto Plc	Basic Materials
ZCC	ZCCM Investments Holdings Plc	Basic Materials
AAL	Anglo American Plc	Basic Materials
GLEN	Glencore Plc	Basic Materials
DGE	Diageo Plc	Consumer Goods
KNM	Konami Holdings Corporation	Consumer Goods
PSN	Persimmon Plc	Consumer Goods
TYT	Toyota Motor Corporation	Consumer Goods
BVIC	Britvic Plc	Consumer Goods
GAW	Games Workshop Group Plc	Consumer Goods
GNC	Greencore Group Plc	Consumer Goods
IMB	Imperial Brands Plc	Consumer Goods
RDW	Redrow Plc	Consumer Goods
ULVR	Unilever Plc	Consumer Goods
BMY	Bloomsbury Publishing Plc	Consumer Services
DEB	Debenhams Plc	Consumer Services
GMD	Game Digital Plc	Consumer Services
HFD	Halfords Group Plc	Consumer Services
MRW	Morrison (Wm) Supermarkets Plc	Consumer Services
TSCO	Tesco Plc	Consumer Services
AO.	AO World Plc	Consumer Services
CFYN	Caffyns Plc	Consumer Services
CCL	Carnival Plc	Consumer Services
CINE	Cineworld Group Plc	Consumer Services
FCCN	French Connection Group Plc	Consumer Services
MONY	Moneysupermarket.Com Group Plc	Consumer Services
PETS	Pets At Home Group Plc	Consumer Services
ADM	Admiral Group Plc	Financials
BARC	Barclays Plc	Financials
HSBA	HSBC Holdings Plc	Financials
SVS	Savills Plc	Financials
UAI	U And I Group Plc	Financials
RBS	Royal Bank Of Scotland Group Plc	Financials
ATMA	Atlas Mara Limited	Financials
BNC	Banco Santander S.A.	Financials
CAY	Charles Stanley Group Plc	Financials
GRI	Grainger Plc	Financials
MTRO	Metro Bank Plc	Financials
GNS	Genus Plc	Health Care
GSK	Glaxosmithkline Plc	Health Care
SHP	Shire Plc	Health Care
PRTC	Puretech Health Plc	Health Care
BTG	BTG Plc	Health Care
AZN	Astrazeneca Plc	Health Care
MDC	Mediclinic International Plc	Health Care
NMC	Nmc Health Plc	Health Care
DPH	Dechra Pharmaceuticals Plc	Health Care
SN.	Smith & Nephew Plc	Health Care
HIK	Hikma Pharmaceuticals Plc	Health Care
BBYB	Balfour Beatty Plc	Industrials
ECM	Electrocomponents Plc	Industrials
GEC	General Electric Company	Industrials
KLR	Keller Group Plc	Industrials
RR.	Rolls-Royce Holdings Plc	Industrials
RMG	Royal Mail Plc	Industrials
AGK	Aggreko Plc	Industrials
CLLN	Carillion Plc	Industrials
ECEL	Eurocell Plc	Industrials
IMI	IMI Plc	Industrials
MTO	Mitie Group Plc	Industrials
BP.	BP Plc	Oil & Gas
PMO	Premier Oil Plc	Oil & Gas
TTA	Total S.A.	Oil & Gas
WG.	Wood Group (John) Plc	Oil & Gas
COPL	Canadian Overseas Petroleum Limited	Oil & Gas
LKOH	PJSC Lukoil	Oil & Gas
CNE	Cairn Energy Plc	Oil & Gas

(*continued on next page*)

Table B2 (*continued*)

Company Ticker	Company Name	Company Industry
XPL	Xplorer Plc	Oil & Gas
TLW	Tullow Oil Plc	Oil & Gas
AVV	Aveva Group Plc	Technology
IBM	International Business Machines Corporation	Technology
SGE	Sage Group Plc	Technology
SDL	SDL Plc	Technology
SCT	Softcat Plc	Technology
USY	Unisys Corporation	Technology
CCC	Computacenter Plc	Technology
FDM	FDM Group (Holdings) Plc	Technology
NCC	NCC Group Plc	Technology
SOPH	Sophos Group Plc	Technology
TOOP	Toople Plc	Technology
KNOS	Kainos Group Plc	Technology
NANO	Nanoco Group Plc	Technology
RM.	RM Plc	Technology
SPT	Spirent Communications Plc	Technology
BT.A	BT Group Plc	Telecommunications
KCOM	KCOM Group Plc	Telecommunications
TDE	Telefonica Sa	Telecommunications
VOD	Vodafone Group Plc	Telecommunications
ISAT	Inmarsat Plc	Telecommunications
TALK	Talktalk Telecom Group Plc	Telecommunications
TEP	Telecom Plus	Telecommunications
CNA	Centrica Plc	Utilities
SVT	Severn Trent Plc	Utilities
UU.	United Utilities Group Plc	Utilities
DRX	Drax Group Plc	Utilities
PNN	Pennon Group Plc	Utilities

References

Aghdam, M. H., Ghasem-Aghaee, N., & Basiri, M. E. (2009). Text feature selection using ant colony optimization. *Expert Systems with Applications, 36*(3), 6843–6853. https://doi.org/10.1016/j.eswa.2008.08.022

Alrubaian, M., Al-Qurishi, M., Alamri, A., Al-Rakhami, M., Mehedi Hassan, M., Fortino, G., & Hassan, M. M. (2018). Credibility in online social networks: A survey. *IEEE Access, 7*, 2828–2855. https://doi.org/10.1109/ACCESS.2018.2886314

Ambusaidi, M., He, X., Nanda, P., & Tan, Z. (2016). Building an intrusion detection system using a filter-based feature selection algorithm. *IEEE Transactions on Computers, 65*(10), 2986–2998. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7387736.

Arauzo-Azofra, A., Aznarte, J. L., & Benítez, J. M. (2011). Empirical study of feature selection methods based on individual feature evaluation for classification problems. *Expert Systems with Applications, 38*(7), 8170–8177. https://doi.org/10.1016/j.eswa.2010.12.160

Ballouli, R. El, El-Hajj, W., Ghandour, A., Elbassuoni, S., Hajj, H., Shaban, K., & Fourier-Grenoble, J. (2017). CAT: Credibility Analysis of Arabic Content on Twitter. Proceedings of the Third Arabic Natural Language Processing Workshop, 62–71. http://shamela.ws/.

Bhattacharya, S., Tran, H., Srinivasan, P., & Suls, J. (2012). Belief surveillance with twitter. Proceedings of the 4th Annual ACM Web Science Conference, WebSci'12, volume, 43–46. https://doi.org/10.1145/2380718.2380724.

Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., & Lang, M. (2020). Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics & Data Analysis, 143*, 106839. https://doi.org/10.1016/j.csda.2019.106839

Bountouridis, D., Sullivan, E., & Hauff, C. (2019). Annotating Credibility : Identifying and Mitigating Bias in Credibility Datasets. ROME 2019 - Workshop on Reducing Online Misinformation Exposure. www.snopes.com.

Castillo, C., Mendoza, M., & Poblete, B. (2013). Predicting information credibility in time-sensitive social media. Internet Research, 23(5), 560–588. https://doi.org/10.1108/IntR-05-2012-0095.

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web* (pp. 675–684).

Castillo, S., Allende-Cid, H., Palma, W., Alfaro, R., Ramos, H. S., Gonzalez, C., Elortegui, C., & Santander, P. (2019). Detection of Bots and Cyborgs in Twitter: A Study on the Chilean Presidential Election in 2017. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11578 LNCS, 311–323. https://doi.org/10.1007/978-3-030-21902-4_22.

Ceccarelli, D., Nidito, F., & Osborne, M. (2016). Ranking financial tweets. SIGIR 2016 - Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, 527–528. https://doi.org/10.1145/2911451.2926727.

Coelho, L., & Richert, W. (2015). Building Machine Learning Systems with Python (2nd ed.). Packt Publishing.

Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2015). Fame for sale: Efficient detection of fake Twitter followers. *Decision Support Systems, 80*, 56–71. https://doi.org/10.1016/j.dss.2015.09.003

Cresci, S., Fabrizio Lillo, Regoli, D., Tardelli, S., Tesoni, M., Lillo, F., Regoli, D., Tardelli, S., & Tesconi, M. (2018). Cashtag piggybacking: uncovering spam and bot activity in stock microblogs on Twitter. ACM Transactions on the Web, 1–18. http://arxiv.org/abs/1804.04406.

Da Cruz, F. M., & De Filgueiras Gomes, M. Y. F. S. (2013). The influence of rumors in the stock market: A case study with Petrobras. Transinformacao, 25(3), 187–193. https://doi.org/10.1590/S0103-37862013000300001.

De Franco, G., Lu, H., & Vasvari, F. P. (2007). Wealth transfer effects of analysts' misleading behavior. Journal of Accounting Research, 45(1), 71–110. https://doi.org/10.1111/j.1475-679X.2007.00228.x.

de Marcellis-Warin, N., Sanger, W., & Warin, T. (2017). A network analysis of financial conversations on Twitter. *Sangerw. Com, 13*(3), 281–309.

De Micheli, C., & Stroppa, A. (2013). Twitter and the underground market. 11th Nexa Lunch Seminar, 5–9. https://nexa.polito.it/nexacenterfiles/lunch-11-de_micheli-stroppa.pdf.

Dorado, H., Cobos, C., Torres-Jimenez, J., Burra, D. D., Mendoza, M., & Jimenez, D. (2019). Wrapper for building classification models using covering arrays. *IEEE Access, 7*, 148297–148312. https://doi.org/10.1109/ACCESS.2019.2944641

Evans, L., Owda, M., Crockett, K., & Vilas, A. F. (2019). A methodology for the resolution of cashtag collisions on Twitter – A natural language processing & data fusion approach. *Expert Systems with Applications, 127*, 353–369. https://doi.org/10.1016/j.eswa.2019.03.019

Gregoriou, G. N. (2015). Handbook of High Frequency Trading. In Handbook of High Frequency Trading. Academic Press. https://doi.org/10.1016/C2014-0-01732-7.

Gupta, A., & Kumaraguru, P. (2012). Credibility ranking of tweets during high impact events. *ACM International Conference Proceeding Series, 10*(1145/2185354), 2185356.

Gupta, A., Kumaraguru, P., Castillo, C., & Meier, P. (2014). Tweetcred: Real-time credibility assessment of content on twitter. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8851, 228–243. https://doi.org/10.1007/978-3-319-13734-6_16.

Hassan, N. Y., Gomaa, W. H., Khoriba, G. A., & Haggag, M. H. (2018). Supervised Learning Approach for Twitter Credibility Detection. Proceedings - 2018 13th International Conference on Computer Engineering and Systems, ICCES 2018, 196–201. https://doi.org/10.1109/ICCES.2018.8639315.

Houlihan, P., & Creamer, G. G. (2019). Leveraging social media to predict continuation and reversal in asset prices. *Computational Economics, 1–21*. https://doi.org/10.1007/s10614-019-09932-9

Hsueh, P.-Y., Melville, P., & Sindhwani, V. (2009). Data Quality from Crowdsourcing: A Study of Annotation Selection Criteria. Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing, 27–35.

Krzysztof, L., Jacek, S.-W., Michal, J.-L., & Amit, G. (2015). Automated Credibility Assessment on Twitter. Computer Science, 16(2), 157. https://doi.org/10.7494/csci.2015.16.2.157.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159. https://doi.org/10.2307/2529310

Liu, Z., Liu, L.u., & Li, H. (2012). Determinants of information retweeting in microblogging. *Internet Research, 22*(4), 443–466. https://doi.org/10.1108/10662241211250980

Loughran, T., & Mcdonald, B. (2016). Textual Analysis in Accounting and Finance: A Survey. Journal of Accounting Research, 54(4), 1187–1230. https://doi.org/10.1111/1475-679X.12123.

Loughran, T., Mcdonald, B., Battalio, R., Easton, P., Fuehrmeyer, J., Gao, P., Harvey, C., Hirschey, N., Marietta-Westberg, J., & Schultz, P. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. The Journal of Finance, 66(1), 35–64. https://www.uts.edu.au/sites/default/files/ADG_Cons2015_Loughran McDonald JE 2011.pdf.

Maddock, J., Starbird, K., Al-Hassani, H., Sandoval, D. E., Orand, M., & Mason, R. M. (2015). Characterizing online rumoring behavior using multi-dimensional signatures. CSCW 2015 - Proceedings of the 2015 ACM International Conference on Computer-Supported Cooperative Work and Social Computing, 228–241. https://doi.org/10.1145/2675133.2675280.

Morris, M. R., Counts, S., Roseway, A., Hoff, A., & Schwarz, J. (2012). Tweeting is believing? Understanding microblog credibility perceptions. Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW, 441–450. https://doi.org/10.1145/2145204.2145274.

Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications, 42*(24), 9603–9611. https://doi.org/10.1016/j.eswa.2015.07.052

Odonovan, J., Kang, B., Meyer, G., Hollerer, T., & Adalii, S. (2012). Credibility in context: An analysis of feature distributions in twitter. Proceedings - 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust and 2012 ASE/IEEE International Conference on Social Computing, SocialCom/PASSAT 2012, 293–301. https://doi.org/10.1109/SocialCom-PASSAT.2012.128.

Oliveira, N., Cortez, P., & Areal, N. (2016). Stock market sentiment lexicon acquisition using microblogging data and statistical measures. *Decision Support Systems, 85*, 62–73. https://doi.org/10.1016/j.dss.2016.02.013

Oliveira, N., Cortez, P., & Areal, N. (2017). The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems with Applications, 73*, 125–144. https://doi.org/10.1016/j.eswa.2016.12.036

Page, J. T., & Duffy, M. E. (2018). What Does Credibility Look like? Tweets and Walls in U.S. Presidential Candidates' Visual Storytelling. *Journal of Political Marketing, 17*(1), 3–31. https://doi.org/10.1080/15377857.2016.1171819

Parmezan, A. R. S., Lee, H. D., & Wu, F. C. (2017). Metalearning for choosing feature selection algorithms in data mining: Proposal of a new framework. *Expert Systems with Applications, 75*, 1–24. https://doi.org/10.1016/j.eswa.2017.01.013

Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M., & Mozetič, I. (2015). The effects of twitter sentiment on stock price returns. PLoS ONE, 10(9), 1–21. https://doi.org/10.1371/journal.pone.0138441.

Rani, D. S., Rani, T. S., & Durga Bhavani, S. (2015). Feature subset selection using consensus clustering. In *ICAPR 2015–2015 8th International Conference on Advances in Pattern Recognition*. https://doi.org/10.1109/ICAPR.2015.7050659

Reidsma, D., & op den Akker, R. (2008). Exploiting "Subjective" Annotations. Workshop on Human Judgements in Computational Linguistics, 8–16.

Ronaghan, S. (2018). The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark. https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3.

Rong, Miao, Gong, Dunwei, & Gao, Xiaozhi (2019). Feature selection and its use in big data: Challenges, methods, and trends. *IEEE Access, 7*, 19709–19725. https://doi.org/10.1109/ACCESS.2019.2894366

Saguna, Zaslavsky, A., & Paris, C. (2012). Context-aware twitter validator (CATVal): A system to validate credibility and authenticity of twitter content for use in decision support systems. Frontiers in Artificial Intelligence and Applications, 238, 323–334. https://doi.org/10.3233/978-1-61499-073-4-323.

Sikdar, S., Adali, S., Amin, M., Abdelzaher, T., Chan, K., Cho, J. H., … O'Donovan, J. (2014). Finding true and credible information on Twitter. In *FUSION 2014–17th International Conference on Information Fusion* (pp. 1–8).

Sikdar, Sujoy, Kang, B., O'donovan, J., Höllerer, T., & Adal, S. (2013). Understanding Information Credibility on Twitter. 2013 International Conference on Social Computing, 19–24. http://www.cs.rpi.edu/~sikdas/papers/socialcom2013.pdf.

Sikdar, Sujoy, Kang, B., O'donovan, J., & Höllerer, T. H. (2013). Cutting Through the Noise: Defining Ground Truth in Information Credibility on Twitter. Human, 2(3), 151–167. https://www.researchgate.net/publication/257200399.

Stringhini, G., Wang, G., Egele, M., Kruegel, C., Vigna, G., Zheng, H., & Zhao, B. Y. (2013). Follow the green: Growth and dynamics in Twitter follower markets. In *Proceedings of the ACM SIGCOMM Internet Measurement Conference* (pp. 163–176). https://doi.org/10.1145/2504730.2504731

Thomson, R., Ito, N., Suda, H., Lin, F., Liu, Y., Hayasaka, R., … Wang, Z. (2012). Trusting tweets: The Fukushima disaster and information source credibility on Twitter. In *ISCRAM 2012 Conference Proceedings - 9th International Conference on Information Systems for Crisis Response and Management* (pp. 1–10).

Tsai, Chih-Fong, & Chen, Yu-Chi (2019). The optimal combination of feature selection and data discretization: An empirical study. *Information Sciences, 505*, 282–293. https://doi.org/10.1016/j.ins.2019.07.091

Yang, F., Liu, Y., Yu, X., & Yang, M. (2012). Automatic Detection of Rumor on Sina Weibo Categories and Subject Descriptors. *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics, 2*.

Yang, J., Yu, M., Qin, H., Lu, M., & Yang, C. (2019). A Twitter data credibility framework - hurricane harvey as a use case. *ISPRS International Journal of Geo-Information, 8*(3), 1–21. https://doi.org/10.3390/ijgi8030111

Yang, Min-Chul, & Rim, Hae-Chang (2014). Identifying interesting Twitter contents using topical analysis. *Expert Systems with Applications, 41*(9), 4330–4336. https://doi.org/10.1016/j.eswa.2013.12.051

Yang, Zhao, Sun, Xuezheng, & Hardin, James W. (2011). Testing marginal homogeneity in clustered matched-pair data. *Journal of Statistical Planning and Inference, 141*(3), 1313–1318. https://doi.org/10.1016/j.jspi.2010.10.002